AUTOMATIC ANNOTATION OF NON-ENGLISH WEB CONTENT

Jakub ŠEVCECH, Mária BIELIKOVÁ

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
{sevcech,bielik}@fiit.stuba.sk

Abstract. It often happens that while reading, we find a word we do not understand. We would welcome an explanation or additional information about this word. For this purpose, annotations in form of additional information retrieved from various information sources are created and attached to such words. In this paper we propose a method for automatic extending the content available on the Web by adding annotations to selected terms (keywords) in the text. The method is designed to be able to insert annotations into the text written in Slovak with a potential to be language independent. Annotations themselves are obtained through publicly available services providing various forms of additional information. We personalize created annotations taking into account implicit feedback from users in form of clickthrough data. We evaluate proposed method in the environment of an educational web-based system.

1. Introduction

While reading a web page, the visitor often encounters a word or phrase, she does not understand, or would require some additional information about this term. This situation occurs more frequently if the page contains technical or explanatory text, such as texts in digital libraries and various educational sites. When a visitor encounters such word, she has to stop work with the document and she has to move attention to other information sources in order to search for required information. One of the solutions to this situation would be an annotation attached to this word. Such annotation can immediately provide explanation of unknown word or additional information that can enrich the context of studied document.

We proposed a method for document content enrichment by attaching additional information to important words of the document. Additional information is retrieved using publicly available services such as search engines or dictionaries. The method attaches annotations to text written in Slovak language, but it has potential to be language independent. We implemented proposed method as web service and we evaluated this method using module in education system ALEF [3].

2. Method for document content enrichment

The described method [1] is composed from several parts:

1. Search for candidate words for annotation attachment
2. Search for additional information to fill the annotation
3. Annotation adaptation and visualization

In the step of searching for candidate words to attach annotation to, we are searching for words for which user could require additional information. This could be specific words (least frequent words in the document corpus), keywords, names of persons, cities, companies or in other way important words. To extract these words, it is possible to use one of many methods of natural language processing such as named entity extraction, keyword extraction etc. In our experiments, we used keywords extracted from the document using keyword and named entity extraction service AlchemyAPI[1]. Machine analysis of text currently achieves satisfactory results only for English texts. We believe that this is sufficient for several languages including Slovak language and therefore we automatically translate analyzed text into the English. Without the use of these results, the quality of found keywords (or other important words) and hence the quality of annotations would decrease significantly. To connect extracted candidate words to their equivalents in the original text, we proposed a method for finding mappings between parallel text translations based on dictionaries and comparison of words using Levenshtein distance.

To search for additional information to fill the annotation, we used multiple services such as Google Search or Slideshare. These services provided us various forms of additional information such as related web page links, definitions, images or slideshows.

The annotation adaptation step uses data about user interaction with created annotations. We use implicit feedback in form of clickthrough data collected when user clicked on presented annotation element and he did not click on the other. We considered these implicit feedbacks to be statements about retrieved annotation content quality and we used them to order these information pieces by decreasing quality. We used several patterns in user clickthrough behavior [2] to extract statements about annotation content quality from users clicks:

1. Click - Skip above: Element user clicked is better than all the elements listed on higher positions and which user did not click
2. Click - No click on next: Element user clicked is better than the immediately following element which user did not click.

We used these statements to create graph, where annotation elements are represented by nodes and their quality statements are represented by directed edges. We used adapted PageRank algorithm to determine annotation element rank. We considered extracted rank

[1] AlchemyAPI, http://www.alchemyapi.com/

to be indicator of annotation element quality and relevance to the document content. We used it to display annotation elements in decreasing order to annotation users.

We performed partial evaluation to determine the precision of candidate words mapping as well as evaluation of entire proposed method in education system Alef. Created annotations provided valuable additional content for important words in documents. The annotation adaptation process was able to order annotation elements according to their quality and relevance to annotated document.

Acknowledgement: This work was partially supported by the Scientific Grant Agency of the Ministry of Education of Slovak Republic, grants VG1/0675/11, KEGA 028-025STU-4/2010 and by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

to other papers publishing the results that are summarized here

[1] Ševcech, J., Bieliková, M.: Automatic Annotation of Non English Web Content. *In: Proc. of the International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, (2011), pp. 281–283.

Other references

[2] Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G.: Accurately interpreting clickthrough data as implicit feedback. *In Proc. of the 28th annual int. ACM SIGIR conf. New York*, ACM, (2005), pp. 154-161.

[3] Šimko, M., Barla, M., and Bieliková, M.: ALEF: A framework for adaptive Web-Based learning 2.0. *In Key Competencies in the Knowledge Society, ser. IFIP Advances in Inf. and Communication Technology*, Springer, (2010), vol. 324, ch. 36, pp. 367-378.